基于内存计算的钢铁价格预测算法研究

朱靖翔 张 滨 乐嘉锦
(东华大学计算机科学与技术学院 上海 201620)

摘 要 由于钢铁价格具有非线性和因子难以确定的特点,在数据挖掘预测分析时,传统的预测方法只能对钢铁价格进行小数据量的分析,这将导致预测精度低、速度慢、效率低下。随着大数据的深入研究,内存计算技术成为研究热点,用户对实时数据处理技术的需求越来越大。因此,在钢铁价格预测模型中,引入内存计算技术,提出基于内存计算的LM-BP神经网络预测算法,利用2002年到2010年的钢铁价格、产量、库存、GDP等数据建立预测模型。最后,仿真实验结果表明,基于内存计算的预测模型算法不仅速度快,而且精度高。

关键词 大数据,内存计算,贝叶斯,ARMA,神经网络
中图法分类号 TP311 **文献标识码** A

Research on Prediction Algorithm Based on In-memory Computing for Steel Prices

ZHU Jing-xiang ZHANG Bin LE Jia-jin
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract Because the steel price is nonlinear and its factor is difficult to be determined, in the forecast analysis, the tranditional method can only analyze the steel price with small amount of data, which leads to low accuracy of predection and the slow speed. In the big data era, memory computing in recent years has been a reseach hotpoint, and the requrement for timely data processing gets larger and larger. Based on the memory computing, the steel prices, production, inventory, and GDP data from 2002 to 2010, were used to build the prediction model, Bayessan forcasting model, ARMA model, support vector machine model and BP neural network model to forecast the steel prices. The simulation results show that the prediction model based on the memory not only has fast speed and high accuracy, but also shows the prices real timely. It provides a strong basis for enterprises to make decision on market reaction.

Keywords Big data, In-memory computing, Bayes, ARMA, Neural networks

钢铁生产在国民经济中具有举足轻重的地位,是社会发展的重要物质保障。中国近50年来在钢铁行业有着突飞猛进的发展。钢铁价格的波动对国民经济以及相关行业带来深远的影响。导致钢铁价格波动的因素有很多,既有来自宏观经济的影响,又有生产成本、供求关系、国际贸易、国家政策等诸多因素。因此,钢铁价格的预测是当前国内外数据挖掘、机器学习应用研究的热点。目前,对钢铁价格预测研究的算法主要有线性回归法、小波分解算法、支持向量机法、聚类分析法等。

在大数据时代,一个好的预测方法可以为决策者提供强有力的决策依据。其不仅可以准确地预测价格的走势,还可以预测顾客对商品的需求,从而让生产者有效地安排生产、减少企业的库存、合理安排物流、提高企业的生产效率、提高顾客对企业的满意程度、综合提高企业的竞争力。目前用于预测的方法有很多,主要的预测方法有:移动平均预测法、指数平滑预测法、趋势外推预测法、回归预测法、灰色预测法、移动自回归预测法(ARIMA)[4]、机器学习法[5]。但是目前的研究工作比较局限于小数据量的预测,在大数据时代,如何有效地利用大数据进行预测有极其重要的意义,这样不仅使得原始数据更有说服力,而且可以使得预测更加精准。重要的是,如果运算的速度足够快,那么我们就可以使预测的即时数据进行实时展现,以为决策者提供最有力的数据支撑,从而大大提高生产效率。

1 相关工作

1.1 内存计算

数据库奠基人Jim Gray曾于2006年预言:"磁带已经死了,磁盘已经落伍,闪存成为新存储,内存局部性才是王道",随着硬件成本的不断降低,如今这一预言已经成为现实。内存计算在软硬件系统协同配置的环境下,高效地将数据库以及数据仓库全部放在内存中进行计算,这样有效地减少了磁盘的I/O。内存计算采用了高效的并行计算技术以及基于内存的数据的读取、处理以及压缩技术,同时支持数据的行式存储以及列式存储。内存计算利用虚拟的数据进行建模,在内存中直接调用有效数据进行分析,减少了数据的冗余,采用系统内置的计算引擎,直接在内存中进行计算,优化了应用层和

本文受"核高基"国家科技重大专项(2010ZX01042-001-003-004)资助。
朱靖翔(1990-),男,硕士生,主要研究方向为数据挖掘、数据库,E-mail:chelsea90@qq.com;张 滨(1978-),男,博士生,讲师,主要研究方向为数据库、数据工程;乐嘉锦(1951-),男,教授,主要研究方向为数据库、数据工程。

数据层之间的数据交互，大大提高了系统的效率。内存计算使得数据的计算速度呈几何级的增长，从而使对海量数据进行实时的分析成为可能[7]。

HANA(*High-Performance Analytic Appliance*)是SAP公司开发的一款内存(in-memory)数据库管理系统。HANA不是一个数据仓库，而是一个基于"列式存储"及"内存计算技术"的软硬件结合体的工作平台[1]。其通过高效的列式存储方式提高数据的压缩效率和存储性能，减少了数据的冗余[2]。用户可以在HANA平台上运用内置的分析工具构建各种分析模型，比如构建数据仓库、报表、仪表盘等，通过直接处理在服务器主存储器上的大量实时数据获得分析和交易的即时结果，从而实现对海量数据的实时分析。HANA的内存数据库(SAP In-Memory Database，IMDB)是其重要组成部分，包括数据库服务器(In-Memory Database Server)、建模工具(Studio)和客户端工具(ODBO、JDBC、ODBC、SQLDBC等)。HANA的计算引擎(Computing Engine)是其核心，负责解析并处理对大量数据的各类CRUDQ操作，支持SQL和MDX语句、SAP和non-SAP数据。SAP HANA非凡的可扩展性和强大的并行计算能力，使得原来需要长时间运行的运营和分析报表将最大程度地实时展现，不再干扰和阻碍其他用户的操作，这将大幅提高企业的生产力和业务效率[3]。

1.2 预测模型与算法

目前用于预测的方法有很多，主要的预测方法有：(1)Bayes预测；(2)时间序列；(3)神经网络模型。

1.2.1 Bayes

贝叶斯预测模型是运用贝叶斯统计进行的一种预测。贝叶斯统计不同于一般的统计方法，其不仅利用了模型信息和数据信息，而且充分利用了先验信息。托马斯·贝叶斯(Thomas Bayes)的统计预测方法是一种以动态模型为研究对象的时间序列预测方法。在做统计推断时，一般模式是：

先验信息＋总体分布信息＋样本信息→后验分布信息

可以看出贝叶斯模型不仅利用了前期的数据信息，还加入了决策者的经验和判断等信息，并将客观因素和主观因素结合起来，对异常情况的发生具有较大的灵活性。

1.2.2 时间序列

时间序列是以时间顺序记录的一系列数据。因为时间单位不同，在一年中记录的数据频次不同。通常以年、季度、月记录的数据，在一年的时间里出现的次数不多，成为低频数据；以周、日等记录的数据为高频数据；日内记录的数据，如分钟或小时，为超高频数据。对不同类型的数据探讨其规律时，采用的方法和模型不尽相同。一般来说，时间序列可以写成下面的形式：

数据＝模型＋误差

ARMA模型

ARMA模型是由美国学者博克斯(George Box)和英国统计学家詹金斯(Gwilym Jenkins)共同建立的，简称B-J法，是一种随机时间序列预测方法。它将预测对象随时间变化形成的序列看做一个随机序列。也就是说，除纯偶然原因引起的个别序列值外，时间序列是依赖于时间 t 的一组随机变量。其中，单个序列值的出现具有不确定性，但是整个序列的变化呈现一定的规律性。B-J方法的基本思想是，这一串随时间变化而又相互关联的数字序列，可以用相应的数学模型近似描述。通过对相应数学模型的分析研究，能更本质地认识这些动态数据的内在结构和复杂特性，从而达到最小方差意义下的最佳预测。

ARMA模型有3种基本类型：自回归(Auto-regressive，AR)模型、移动平均(Moving Average，MA)模型以及自回归移动平均(Auto-regressive Moving Average，ARMA)模型。

1.2.3 BP神经网络

BP网络是一种多层前馈型神经网络，其神经元的传递是S型函数，输出量为0到1之间的连续量，它可以实现从输入函数到输出的任意非线性映射。由于权值的调整采用反响传播学习方法，因此也常称其为BP网络(Back Propagation Network)[11]。图1为BP神经元模型。

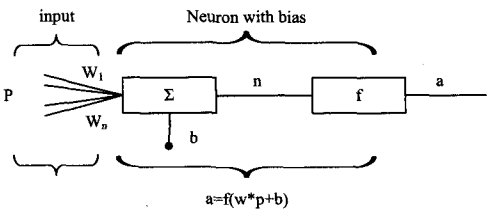

图1 BP神经元模型

2 基于内存计算的LM-BP神经网络预测算法

钢铁价格由于存在很强的不确定性，既受宏观层面的影响，比如国家的政策、全球经济大环境、生产的成本、资源状况等，也受到微观层面的影响，比如供求关系、企业调价、库存状况等因素。综合宏观层面以及微观层面的影响，结合现有的数据，基于内存计算的钢铁价格预测算法选出人民币美元汇率、国民生产总值(GDP)、进口总量、国民消费总值、铁矿石成本价，这5个影响钢铁价格的主要因素作为LM-BP神经网络模型的输入层进行分析预测。

2.1 隐含层节点数的选择

隐含层节点数直接影响网络结构模型效果的好坏。因此，在BP网络中，隐含层节点数的选择非常重要。如果隐含层的结点数太少，网络模型则不能达到必要的学习效果和预测的精度[8]。如果隐含层节点数过多，不仅会大大增加网络结构的复杂性，网络在学习过程中更易陷入局部极小点，而且会使网络模型的训练速度变慢[9]。为了尽可能地避免网络模型在训练时出现"过拟合"的现象，保证网络模型有着足够高的精度和泛化能力，确定隐含层节点数的最基本原则是：在满足精度要求的前提下取尽可能少的隐含节点数[6]。

Gorman定理：$S=\log_2 N$(S：隐含结点数，N：模式数)；

Kolmogorov定理：$S=2n+1$(S：隐含结点数，n：输入层结点数)。

输入层节点数为5，我们选取隐含层节点数为11。

2.2 数据的标准化

由于该预测模型中输入层为国民生产总值、进口总量、库存量和国民消费总值、铁矿石成本价，其单位分别是亿元、万吨、万吨、亿元和元，其计量单位以及数量级都不相同，用原始数据作为输入向量会有较大数值差别的特征值，会使网络模型难以收敛，从而严重影响预测的效果。因此我们有必要对输入的数据进行归一化处理。标准化方法是对原始数据进行线性变换。设 A_{min} 和 A_{max} 分别为属性 X 的最小值和最大值，将 X 的一个原始值 X 通过标准化映射成在区间[0，1]中的值 Y，其公式为：

$$Y=(X-A_{min})/(A_{max}-A_{min}) \quad (1)$$

式中，X 为原始数据，Y 为归一化以后的数据。

至此，BP 神经网络预测模型确定为 5×11×1 的结构。按照 BP 网络的一般设计原则，中间层神经元的传递函数为 S 型正切函数。由于输出已经被归一化到区间[0,1]中，因此，输入层神经元的传递函数可以设定为 S 型对数函数。

S 型函数具有非线性放大系数功能，可以把输入从负无穷大到正无穷大的信号，变换成－1 到 1 之间的输出。对较大的输入信号，放大系数较小；而对较小的输入信号，放大系数则较大。采用 S 型激活函数可以处理和逼近非线性输入/输出关系。

2.3 基于内存计算的 LM 算法

标准的 BP 神经网络的算法有着收敛速度慢、学习速率不易确定的缺点。目前常用的优化算法有变速率算法、附加动量算法、共轭梯度算法、高斯-牛顿算法、Levenberg-Marquardt 算法(简称 LM 算法)等，其中 LM 算法是这些算法中收敛速度最快、鲁棒性最好的[10]。LM 算法能够根据迭代的结果动态地调整阻尼因子来动态地调整迭代的收敛方向，可使每次的迭代误差函数值都有所下降[5]。但是 LM 算法的运算量大，对内存的要求极高，如果在样本的数据量非常大运算非常复杂的情况下，用 LM 算法进行训练，效果很差，所以 LM 算法在以往的实际应用中并没有普及。随着内存计算的发展，HANA 可以很好地克服 LM 算法对内存要求高的问题。因此，本实验中 BP 神经网络我们采用 LM 的算法进行训练。

算法的具体描述：

评价函数 $E(a)=\sum_{m-1}^{n} f_i^2(a)$

1)$\mu<-10^{-1}$，$a(0)=a_0$

2)计算 $\sum(a)$

3)计算 Jacobian 矩阵 $J=[\frac{\partial f}{\partial a_1}]$

4)$a(k+1)=a(k)1[J^TJ+\mu_1]^{-1}J^Tf(a(k))$

5)计算 $E(a(k+1))$

6)如果 $E(a(k+1))>E(a(k))$，则 $\mu\leftarrow\mu\times10$，转到 4)

7)如果 $E(a(k+1))<$误差指标，则计算成功，算法结束

8)$\mu\leftarrow\mu\times0.1$

9)$a(k)=a(k+1)$，转到 2)

自此，BP 神经网络价格预测模型构造完毕。用已经导入到 HANA 中的相关数据作为输入值，不断地训练初始神经网络，通过不断地调整权值以及阈值，来不断地提高预测模型预测的效果，直到训练的精度达到规定的要求。最后将模型所得预测值和准备好的测试数据作比较，检验预测的精度是否达到预期。图 2 为 BP 网络实现流程图。

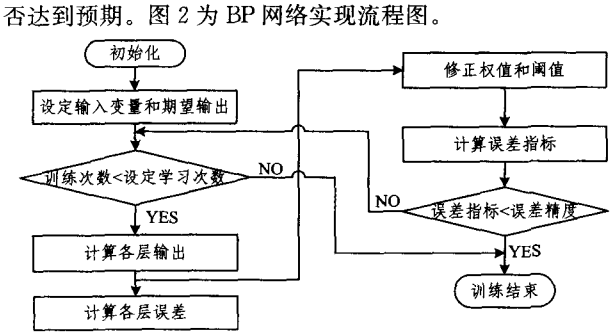

图 2 BP 网络实现流程

算法实现步骤如算法 1 所示。

算法 1 LM-BP 神经网络

输入：汇率 GDP，进口总量，国民消费总值，铁矿石成本价

输出：钢铁预测价格

1. u=0.1，β=10，k=0；

2. $d_o(k)=(d_1(k),d_2(k),\cdots,d_q(k))^\varepsilon$，$x(k)=(x_1(k),x_2(k),\cdots,x_n(k))$；

3. $E(x^k)=\frac{1}{2m}\sum_{k=1}^{m}\sum_{o=1}^{q}(d_o(k)-y_o(k))^2$；

5. Jacobian MATRIX J(e(x^k))；

6. $\Delta x=-[J^K(x)J(x)+\mu I]^{-1}J(x)e(x)$

7. IF $E(x^k)<\varepsilon$，THEN GOTO 9

ELSE

$E(x^{k+1})$；

8. IF $E(x^{k+1})<E(x^k)$，THEN x^{k+1}

$x^{k+1}=x^k+\Delta x$，u=u/β，k=k+1

RETURN 2，

ELSE x^k//即下次迭代的 Jacobian 矩阵不变

u=u·β，k=k+1，

IF k=M THEN GOTO 9

ELSE GOTO 6；

9. END

3 实验

3.1 实验设置

本实验所用服务器的配置为戴尔 PowerEdge R910，CPU 为 4 颗 Xeon 8 核 E7520，内存 256G，操作系统为 SUSE Linux Enterprise Server 11 SP1，内存计算数据库引擎采用 HANA SERVER 1.006。实验数据采用 2002 年 1 月至 2010 年 8 月的热轧钢价格基础数据，如表 1 所列。

表 1 2002 年 1 月至 2010 年 8 月的热轧钢数据

日期	汇率	产量	进口量	GDP	全民消费总值
2002 年 1 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 2 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 3 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 4 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 5 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 6 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 7 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 8 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 9 月	827.67	1292.78	785.00	6693.929739	3596.1000
2002 年 10 月	827.66	1262.22	602.00	6874.665842	3324.4000
2002 年 11 月	827.66	1262.22	602.00	6874.665842	3324.4000
2002 年 12 月	827.66	1262.22	602.00	6874.665842	3324.4000
2003 年 1 月	827.70	1370.29	993.00	7624.004419	3114.8000
…	…	…	…	…	…
2010 年 8 月	827.70	1370.29	993.00	7624.004419	3114.8000

3.2 结果与分析

因为现实中的数据含有噪音和不完整性，需要对其进行处理，检测异常数据。而在 BP 神经网络中一般是需要通过数据的归一化处理，使输入/出数据落在神经元激活函数较大的区域，以提高预测精度和收敛速度。一般是将输入数据限制在[0,1]之间，但是考虑到 0 和 1 都是神经网络中响应函数的上下极限值，不宜作为输入、输出的实际中使用，本文在实际的实验中采用式(1)对数据进行了归一化处理。

选择钢铁价格为研究对象，将 2002 年、2006 年和 2008

年12个月的实际钢铁价格输入LM-BP网络，2005年、2007年和2009年钢铁价格作为训练样本的期望值。训练结果如图3和表2所示。由图3可知，在经过6次迭代之后，数据呈现快速收敛。

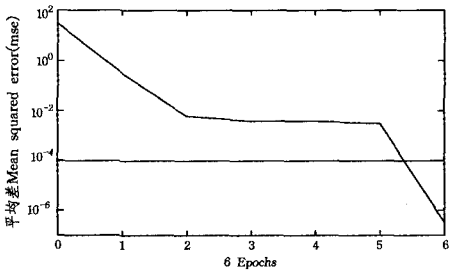

图3 LM-BP网络算法训练示意图

表2 归一化后2002年1月至2010年8月的热轧钢数据

日期	汇率	产量	进口量	GDP	全民消费总值
Jan/02	0.9996	0.00771	0.03127	0.00148	0.05691
Feb/02	0.9996	0.00771	0.03127	0.00148	0.05691
Mar/02	0.9996	0.00771	0.03127	0.00148	0.05691
Apr/02	0.9996	0.00771	0.03127	0.00148	0.05691
May/02	0.9996	0.00771	0.03127	0.00148	0.05691
Jan/02	0.9996	0.00771	0.03127	0.00148	0.05691
Jun/02	0.9996	0.00771	0.03127	0.00148	0.05691
Jul/02	0.9996	0.00771	0.03127	0.00148	0.05691
Aug/02	0.9996	0.00771	0.03127	0.00148	0.05691
Feb/02	0.99953	0	0	0.00565	0.02848
…	…	…	…	…	…
Aug/10	0.9998	0.02726	0.0668	0.02293	0.00655

3.2.1 数据处理性能对比

数据量变化下对算法验证

通过对比分析实验发现，当样本数量增加时，与标准的BP神经网络实验相比较，使用LM算法的BP神经网络执行速度优化提升很明显，精度也有显著提高。在1年的样本数据下，使用LM算法的BP神经网络实验比标准BP神经网络实验在时间上减少27.6%，精度上提升23.4%，在10年的样本数据下，使用LM算法的BP神经网络实验比标准BP神经网络实验在时间上减少37.5%，精准度上提升31.2%。

3.2.2 准确率对比

算法在传统数据库与当前内存数据库HANA中的对比验证预测结果最后100日的平均绝对误差曲线如图4所示。

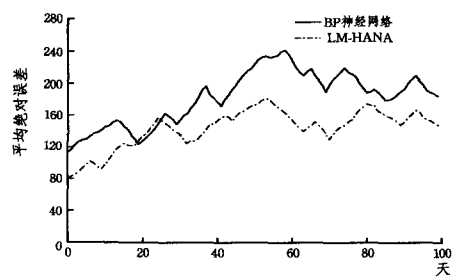

图4 预测结果最后100日的平均绝对误差曲线

通过对比实验分析发现，使用LM算法的LMBP-HANA原型系统执行速度优化明显，同样是10年的钢铁价格数据，在HANA环境下用LM算法进行训练，速度有着明显的优化，很好地克服了LM算法运算复杂、对内存要求高的问题。LMBP-HANA原型系统比DBMS系统平均执行的时间降低43.8%，精度提高34.5%。

结束语 钢铁的价格变化波动不定，传统的预测方法仅仅能够对少量的数据进行预测分析。本文采用了基于内存计算的LM-BP神经网络预测算法，建立了新的预测模型，对2002年到2010年的钢铁的数据进行预测分析。实验结果表明，内存计算很好地解决了该算法运算量大、时间复杂度高的不足，可以有效地预测钢铁的价格走势，不仅比传统的预测方法速度更快，而且精度更高，具有很好的应用前景。

参考文献

[1] Sikka V, Färber F, Lehner W, et al. Efficient transaction processing in SAP HANA database: the end of a column store myth [C] // Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012: 731-742

[2] Rösch P, Dannecker L, Färber F, et al. A storage advisor for hybrid-store databases[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 1748-1758

[3] Färber F, Cha S K, Primsch J, et al. SAP HANA database: data management for modern business applications [J]. SIGMOD Record, 2011, 40(4): 45-51

[4] 邹柏贤，刘强. 基于ARMA模型的网络流量预测[J]. 计算机研究与发展，2002，39(12)：1645-1652

[5] 彭岩，王万森，王旭仁. 基于机器学习的风险预测方法研究[J]. 计算机科学，2009，36(4)：205-207

[6] 贾丽会，张修如. BP算法分析与改进[J]. 计算机技术与发展，2006，16(10)：101-103

[7] 覃雄派，王会举，李芙蓉. 数据管理技术的新格局[J]. 软件学报，2013，24(2)：175-197

[8] 师洪涛，杨静玲，丁茂生. 基于小波-BP神经网络的短期风电功率预测方法[J]. 电力系统自动化，2011，35(16)：44-48

[9] 柳进，于继来，唐降龙. 基于数据挖掘的电网高峰负荷预测系统[J]. 计算机工程，2005，31(1)：9-11

[10] 姚立忠，李太福，易军. 神经网络模型的透明化及输入变量约简[J]. 计算机科学，2012，39(9)：247-251

[11] 韩力群. 人工神经网络教程[M]. 北京：北京邮电大学出版社，2006：12-132

[12] 孙红敏，吴静婷，李晓明. 基于改进BP神经网络的价格预测模型研究[J]. 东北农业大学学报，2013，44(8)：133-137

[13] 韩震，赵宁. 基于LM-BP神经网络的Argo数据西北太平洋海水温度模型[J]. 海洋环境科学，2012，31(4)：555-560

[14] 王卫东，李净，张福存，等. 基于BP神经网络的太阳辐射预[J]. 干旱区资源与环境，2014，28(2)：185-189

[15] 徐黎明，王清，陈剑平，等. 基于BP神经网络的泥石流平均流速预测[J]. 吉林大学学报：地球科学版，2013，43(1)：186-191

[16] 欧阳红祥，李欣，张信娟. 人工神经网络在建筑材料价格预测中的应用[J]. 武汉理工大学学报：信息与管理工程版，2013，35(1)：115-118